Temporal Decision Trees

G. Rizk, Ahlame Douzal Chouakria, Cécile Amblard

Temporal Decision Trees

Guillaume Rizk
TIMC-IMAG-CNRS (UMR 5525), Université Joseph Fourier Grenoble 1,
F-38706 LA TRONCHE Cedex, France
E-mail: Guillaume.Rizk@imag.fr

Ahlame Douzal Chouakria
TIMC-IMAG-CNRS (UMR 5525), Université Joseph Fourier Grenoble 1,
F-38706 LA TRONCHE Cedex, France
E-mail: Ahlame.Douzal@imag.fr

Cécile Amblard
TIMC-IMAG-CNRS (UMR 5525), Université Joseph Fourier Grenoble 1,
F-38706 LA TRONCHE Cedex, France
E-mail: Cécile.Amblard@imag.fr

1 Introduction

Classification problem with time series input variables arise naturally in many applications. In this paper, we propose a new approach for extending standard decision trees (Breiman et al. (1984)) to handle time series input variables. Our proposition aims to leverage the interpretability of the temporal decision tree as well as its accuracy and performance. Past contributions to this problem, can be classified into two main approaches. The first one maps time series to another description space, then perform conventional classifiers on the transformed time series (Geurts (2001)). The second approach works directly on the time series, they generally propose to split a time series variable based on the proximity measure between time series (Yamada et al. (2005)). Although, the direct approach offers more comprehensive and interpretable output results than the former one, it suffers of two limitations. On the one hand, the used proximity measures, the euclidean distance or the dynamic time warping, are based on the closeness of the values regardless to the proximity with respect to the behavior of the time series. On the other hand, proximity measure is only based on the whole time series, ignoring the case of subsequences providing a more optimal split of internal nodes. In this paper we propose a new temporal decision tree which handles directly time series. The main idea of our proposition can be summarized in the following points: using a dissimilarity index including both proximity on values and on behaviors; for the split of an internal node a time series is considered on its whole length or on a subsequence; the research of the optimal split consists in learning the dissimilarity index which minimizes the Gini index with a good separability between the two obtained daughter nodes. The rest of the paper is organized as follows. In the next section, we present the main principals of the dissimilarity index including both behavior and values proximity measures. The section 3 presents the new decision tree method and gives the main algorithms. In section 4, we illustrate the high benefit and performance of our new proposition through several public datasets; we perform its evaluation and comparison to alternative approaches and discuss the main obtained results.

2 Adaptive dissimilarity index for time series classification

We distinguish two important characteristics of the temporal applications. On the one hand, there are applications where both occurring events and their instants of time are determinant for the proximity evaluation. For instance, ECG, delay response to a treatment, etc. We characterize such applications

as "Time dependent events". On the other hand, there are the applications where only occurring events and their order which are important. For instance, in voice processing domain only the occurring syllables are used to identify words; the flow rate being specific to each person. We characterize such applications as "Time independent events". To include both proximity on values and on behavior for time series proximity measure, it's appropriate for the former type of applications to use the extended euclidean distance proposed in (Douzal Chouakria et al. (2007)), where as for the latter type of application, we propose a new extension of the dynamic time warping.

Time dependent events application Let $S_1 = (u_1, ..., u_p)$ and $S_2 = (v_1, ..., v_p)$ be two time series of p values observed at the time instants $(t_1, ..., t_p)$ and assumed as issued from a "Time dependent events" application. The adaptive dissimilarity index proposed in (Douzal Chouakria et al. (2007)) to extend the euclidean distance to include both proximity measures with respect to values and with respect to (w.r.t.) behavior is defined as follows:

$$(1) \qquad D_E^k(S_1, S_2) = f(cort(S_1, S_2)) \cdot \delta_E(S_1, S_2) \qquad \text{with} \qquad f(x) = \frac{2}{1 + exp(k\, x)} \quad , \quad k \geq 0$$

$\delta_E^k(S_1, S_2) = \left(\sum_{i=1}^{p}(u_i - v_i)^2\right)^{\frac{1}{2}}$ is the conventional euclidean distance based on the closeness of the values and $cort(S_1, S_2) = \frac{\sum_{i=1}^{p-1}(u_{(i+1)} - u_i)(v_{(i+1)} - v_i)}{\sqrt{\sum_{i=1}^{p-1}(u_{(i+1)} - u_i)^2}\sqrt{\sum_{i=1}^{p-1}(v_{(i+1)} - v_i)^2}}$ is the temporal correlation coefficient ($cort \in [-1, 1]$) and defines a similarity w.r.t. behavior. The parameter k modulates the contributions of the proximity w.r.t. values and w.r.t. behavior to the dissimilarity index D_E^k. Note that if k varies in [0, 6] $f(x)$ decreases from 2 to 0.

Time independent events application Let's now assume that $S_1 = (u_1, ..., u_p)$ and $S_2 = (v_1, ..., v_q)$ are two time series issued from a "Time independent events" application. We define a mapping $r \in M$ (M is the set of all possible mappings) between S_1 and S_2 as the sequence of m pairs preserving the observations order $r = ((u_{a_1}, v_{b_1}), (u_{a_2}, v_{b_2}), ..., (u_{a_m}, v_{b_m}))$ with $a_i \in \{1, .., p\}$, $b_j \in \{1, ..q\}$ and satisfying for $i \in \{1, .., m-1\}$ the following constraints: $a_1 = 1$, $a_m = p$, $a_{i+1} = (a_i \text{ or } a_i + 1)$ and $b_1 = 1$, $b_m = q$ and $b_{i+1} = (b_i \text{ or } b_i + 1)$. Let's note $S_1^r = (u_{a_1}, ..., u_{a_m})$ and $S_2^r = (v_{b_1}, ..., v_{b_m})$ the two time series induced by such a mapping r. We define the length $|r|$ of such a mapping as:

$$(2) \qquad |r| \quad = \quad f(cort(S_1^r, S_2^r)). \sum_{i=1,..,m} |u_{a_i} - v_{b_i}|$$

According to the above definition, we propose a new adaptive dynamic time warping to cover both proximity on values and on behavior :

$$(3) \qquad D_{dtw}^k(S_1, S_2) \quad = \quad \min_{r \in M} |r| = \min_{r \in M} \left(f(cort(S_1^r, S_2^r)). \sum_{i=1,..,m} |u_{a_i} - v_{b_i}| \right)$$

3 Temporal decision trees

To extend decision trees to time series input variables, we propose a new adaptive split procedure to partition the set of time series into two clusters minimizing the Gini error. If we suppose given a dissimilarity index, and two representative time series, we define a two clusters partition of the set of time series by assigning each time series to the closest (in terms of the given dissimilarity) representative time series. The novelty of our adaptive split procedure is summed up in two points. On the one hand, instead assuming a given dissimilarity index (i.e. the euclidean or the dynamic time warping), we have to learn the best dissimilarity index (i.e. the best contributions of the values and of the behavior) to provide a two clusters partition minimizing the Gini error. On the other hand, we broaden the search of the two representative time series to a dichotomic search of sub time series providing a more optimal partition (in terms of Gini index). Let's give more in detail the proposed algorithms. Let $S = \{1, ..., N\}$

be the set of time series belonging to a current node. The split of a set of time series is characterized by $\sigma(l, r, k, I)$ where l and r identify the left and right representative time series, k the parameter defining the contribution of the behavior and of the values in the dissimilarity index D^k (with D^k referring D^k_E or D^k_{dtw}), and I is the observation period of the studied time series. For a given observation period I, we define two consecutive overlapped sub-periods I_L, I_R subdividing I into a left and a right sub-periods; with $I_L \cup I_R = I$ and $I_L \cap I_R \neq \phi$. Finally let's note $GI(\sigma(l, r, k, I))$ the Gini error of the split $\sigma(l, r, k, I)$.

Algorithm 1 Dichotomic-Split(S,I)

$I_L \leftarrow$ Left sub period of I
$I_R \leftarrow$ Right sub period of I
$(\sigma(l^I_*, r^I_*, k^I_*, I), error_I) \leftarrow$ Adaptive-Split(S,I)
$(\sigma(l^{I_L}_*, r^{I_L}_*, k^{I_L}_*, I_L), error_{I_L} \leftarrow$ Adaptive-Split(S,I_L)
$(\sigma(l^{I_R}_*, r^{I_R}_*, k^{I_R}_*, I_R), error_{I_R}) \leftarrow$ Adaptive-Split(S,I_R)
if $error_{I_L} = \min(error_{I_L}, error_{I_R}, error_I)$ **then**
  $I_* \leftarrow I_L$
else if $error_{I_R} = \min(error_{I_L}, error_{I_R}, error_I)$ **then**
  $I_* \leftarrow I_R$
else
  return $< \sigma(l^I_*, r^I_*, k^I_*, I) >$
end if
$\sigma(l^{I_*}_*, r^{I_*}_*, k^{I_*}_*, I_*) \leftarrow$ Dichotomic-Split(S,I_*)
return $< \sigma(l^{I_*}_*, r^{I_*}_*, k^{I_*}_*, I_*) >$

Algorithm 2 Adaptive-Split(S,I)

$best\_error \leftarrow \infty$
for k in $[0; 6]$ **do**
  $(l_k, r_k) \leftarrow argmin_{(l,r)}(GI(\sigma(l, r, k, I)))$
  if $GI(\sigma(l_k, r_k, k, I)) < best\_error$ **then**
    $best\_error \leftarrow GI(\sigma(l_k, r_k, k, I))$
    $k_* \leftarrow k$
    $l_* \leftarrow l_k$
    $r_* \leftarrow r_k$
  end if
end for
return $< \sigma(l_*, r_*, k_*, I), best\_error >$

Initially, the $Dichotomic - Split$ procedure is called with the interval I corresponding to the total observation period. The $Adaptive - Split$ procedure is then called with I, I_L, and I_R periods. Given as input an observation period I, the $Adaptive - Split$ procedure will look for the best dissimilarity D^k (i.e. the best k) and the best representative time series based on the observation within the period I (i.e. dissimilarity index is evaluated on the sub-sequences defined by I). As output, the $Adaptive - Split$ returns the best parameter k^I_* (the best contribution of the values and of the behavior for the dissimilarity index D^k) and the two representative time series l^I_* and r^I_* providing an optimal split in terms of Gini error. The dichotomic search continue as long as at least one of the left or of the right sub-periods improve the Gini error. As output, the $Dichotomic - Split$ returns the best observation period I_* (which can be the total initial observation period) and the optimal corresponding split $\sigma(l^{I_*}_*, r^{I_*}_*, k^{I_*}_*, I_*)$.

4 Application and comparison results

We illustrate the efficiency of the proposed algorithm through 4 simulated datasets described in (Geurts (2005)): *CBF, CBF-tr, CC* and *Two-Pat.* As these datasets are of "independent time events" type, D^k_{dtw} is considered. Our proposition is compared to the Yamada's algorithm [Yamada et al. (2005)] and to "Segment and Combine (S&C)" procedure (Geurts (2005)). Data are noisy preprocessed. Comparison results are presented in Table 1. For instance, the third row shows that a decision tree built on the CC dataset, which are composed of 900 training cases and 300 test cases, gives a global error rate of 0.003 for our proposition, 0.023 for the Yamada's approach and 0.003 for S&C one. The numbers of leaves are indicated between brackets. Note that through nearly all the datasets, the proposed algorithm provides a better error rate than the two alternative approaches. Let's explain the main characteristics of the decision tree (figure 1) built on the CC data. Initially the root node (node 1) includes 900 times series equally distributed through the 6 CC classes named: *Cyclic*, *Decreasing*, *Downward*, *Increasing*, *Normal* and *Upward.* The split of that node reveals 3 main elements: the two time series (of downward and cyclic classes) selected as the most representative of the left and

	Training set size	Test set size	New tree	Yamada	S&C
CBF	600	300	0.006 (3)	0.066 (3)	0.015
CBF-tr	600	300	0.013 (3)	0.116 (15)	0.027
CC	900	300	0.003 (6)	0.023 (7)	0.003
Two-pat	400	400	0.005 (4)	0.000 (6)	0.049

Table 1: Experiments results (*error rate (Nb. leaves)*)

right daughter nodes, the optimal value $k* = 0$ meaning that, to discriminate well the daughter nodes time series, the dissimilarity index should be based mainly on the values, and finally, the retained discriminant interval I_* (here the whole observation period) localizing the observation period on which the dissimilarity index should be evaluated. Given a new time series case, we evaluate the dissimilarity index D_{dtw}^{k*} (limited to the observations belonging to I_*) between the new case and each of the representative time series; the new time series is then assigned to the daughter node of the most similar representative time series. Finally, let's note that the split of the node 4 reveals a value of $k_* = 5$, which means that the dissimilarity index to be used for assigning new cases should be based essentially on the behavior.

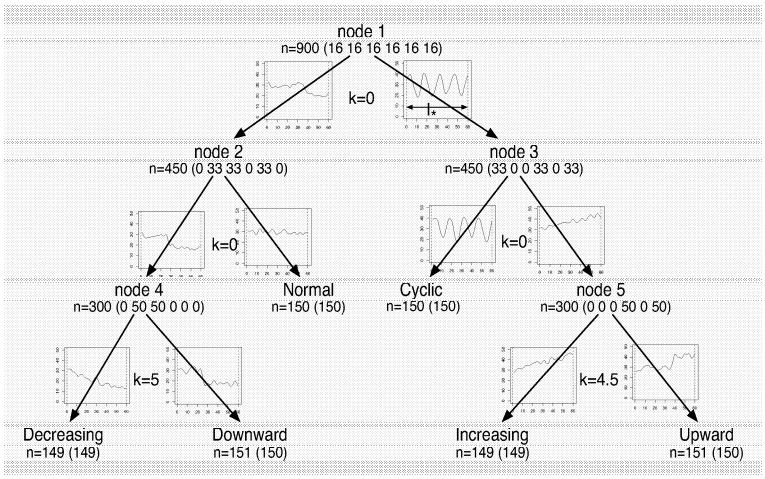

Figure 1: Temporal decision tree for CC datasets classification

5 Conclusion

We propose a new splitting approach to extend the decision trees to temporal data. The proposed split aims to determine for each daughter node the representative time series, the observation period best discriminating the output variable, and the optimal contribution of the values and of the behavior for the proximity evaluation. A new extension of the Dynamic time warping is also proposed. The high efficiency and interpretability of the proposition is illustrated through many public datasets and compared to two important alternative algorithms. Future work will focus on the stability evaluation through real datasets.

REFERENCES

Douzal Chouakria A., Nagabhushan P. (2007), Adaptive dissimilarity index for measuring time series proximity. Advances in Data Analysis and Classification Journal. Vol 1, 5-21 Springer.

Breiman L., Friedman J. et al. (1984), classification and regression trees, Wadsworth, Belmont, CA.

Geurts P. (2005), Contribution to decision tree induction: thesis, 2002

Yamada et al. (2005), Experimental evaluation of time series decision tree, LNCS, vol 3430, p190-209.